PERSPECTIVE

Perspective: NutriGrade: A Scoring System to Assess and Judge the Meta-Evidence of Randomized Controlled Trials and Cohort Studies in Nutrition Research[1–3]

Lukas Schwingshackl,[5]* Sven Knüppel,[5] Carolina Schwedhelm,[5] Georg Hoffmann,[6] Benjamin Missbach,[6] Marta Stelmach-Mardas,[5,7] Stefan Dietrich,[5] Fabian Eichelmann,[4,5] Evangelos Kontopanteils,[8] Khalid Iqbal,[5] Krasimira Aleksandrova,[4,5] Stefan Lorkowski,[9,10] Michael F Leitzmann,[11] Anja Kroke,[12] and Heiner Boeing[5]

[4]Nutrition, Immunity, and Metabolism Start-Up Lab, [5]Department of Epidemiology, German Institute of Human Nutrition Potsdam Rehbruecke, Nuthetal, Germany; [6]Department of Nutritional Sciences, University of Vienna, Vienna, Austria; [7]Department of Pediatric Gastroenterology and Metabolic Diseases, Poznan University of Medical Sciences, Poznan, Poland; [8]Centre for Primary Care, Institute of Population Health, University of Manchester, Manchester, United Kingdom; [9]Institute of Nutrition, Friedrich Schiller University Jena, Jena, Germany; [10]Competence Cluster of Nutrition and Cardiovascular Health, Halle-Jena-Leipzig, Germany; [11]Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany; and [12]Department of Nutritional, Food, and Consumer Sciences, University of Applied Sciences, Fulda, Germany

ABSTRACT

The objective of this study was to develop a scoring system (NutriGrade) to evaluate the quality of evidence of randomized controlled trial (RCT) and cohort study meta-analyses in nutrition research, building upon previous tools and expert recommendations. NutriGrade aims to assess the meta-evidence of an association or effect between different nutrition factors and outcomes, taking into account nutrition research–specific requirements not considered by other tools. In a pretest study, 6 randomly selected meta-analyses investigating diet–disease relations were evaluated with NutriGrade by 5 independent raters. After revision, NutriGrade was applied by the same raters to 30 randomly selected meta-analyses in the same thematic area. The reliability of ratings of NutriGrade items was calculated with the use of a multirater κ, and reliability of the total (summed scores) was calculated with the use of intraclass correlation coefficients (ICCs). The following categories for meta-evidence evaluation were established: high (8–10), moderate (6–7.99), low (4–5.99), and very low (0–3.99). The NutriGrade scoring system (maximum of 10 points) comprises the following items: *1*) risk of bias, study quality, and study limitations, *2*) precision, *3*) heterogeneity, *4*) directness, *5*) publication bias, *6*) funding bias, *7*) study design, *8*) effect size, and *9*) dose-response. The NutriGrade score varied between 2.9 (very low meta-evidence) and 8.8 (high meta-evidence) for meta-analyses of RCTs, and it ranged between 3.1 and 8.8 for meta-analyses of cohort studies. The κ value of the ratings for each scoring item varied from 0.32 (95% CI: 0.22, 0.42) for risk of bias for cohort studies and 0.95 (95% CI: 0.91, 0.99) for study design, with a mean κ of 0.66 (95% CI: 0.53, 0.79). The ICC of the total score was 0.81 (95% CI: 0.69, 0.90). The NutriGrade scoring system showed good agreement and reliability. The initial findings regarding the performance of this newly established scoring system need further evaluation in independent analyses. *Adv Nutr* 2016;7:994–1004.

Keywords: meta-analysis, meta-evidence, nutrition, reliability, scoring

Introduction

Evidence-based dietary recommendations should be based on the completeness of the available evidence (1). Systematic reviews and meta-analyses represent the most valuable, reliable, and objective tool to summarize the evidence for a specific research question in nutrition research (2, 3). However, many meta-analyses have not evaluated the quality of the evidence, decreasing our confidence in the observed effect.

[1] The authors reported no funding received for this study. Perspective articles allow authors to take a position on a topic of current major importance or controversy in the field of nutrition. As such, these articles could include statements based on author opinions or point of view. Opinions expressed in Perspective articles are those of the author and are not attributable to the funder(s) or the sponsor(s) or the publisher, Editor, or Editorial Board of *Advances in Nutrition*. Individuals with different positions of the topic of a Perspective are invited to submit their comments in the form of a Perspectives article or as a Letter to the Editor.

[2] Author disclosures: L Schwingshackl, S Knüppel, C Schwedhelm, G Hoffmann, B Missbach, M Stelmach-Mardas, S Dietrich, F Eichelmann, E Kontopanteils, K Iqbal, K Aleksandrova, S Lorkowski, MF Leitzmann, A Kroke, and H Boeing, no conflicts of interest.

[3] Supplemental Tables 1 and 2, Supplemental Appendix, and Supplemental References are available from the "Online Supporting Material" link in the online posting of the article and from the same link in the online table of contents at http://advances.nutrition.org.

*To whom correspondence should be addressed. E-mail: lukas.schwingshackl@dife.de.

[13] Abbreviations: AMSTAR, assessing the methological quality of systematic reviews; GRADE, Grading of Recommendations Assessment, Development and Evaluation; ICC, intraclass correlation coefficient; RCT, randomized controlled trial.

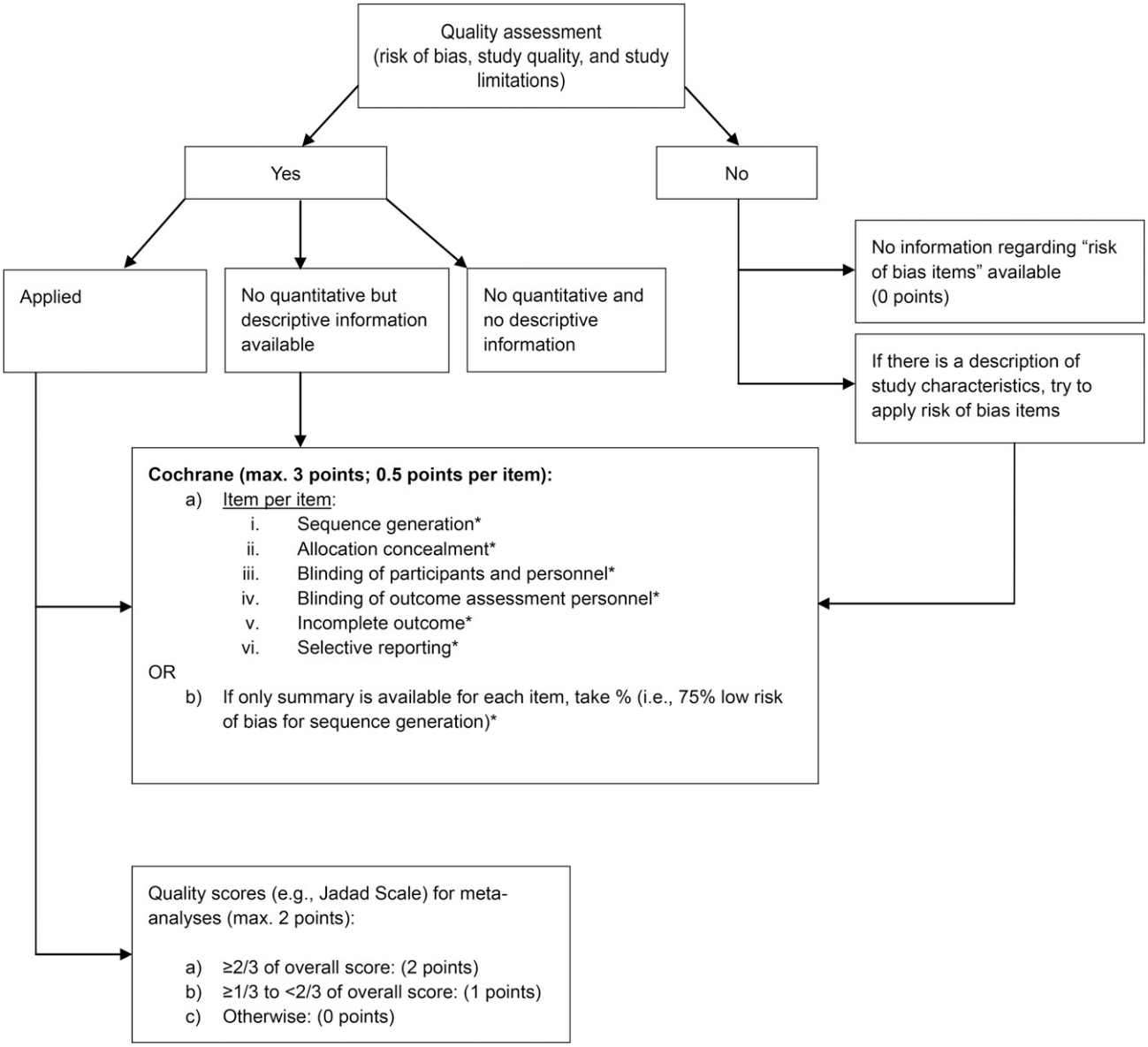

FIGURE 1 Procedure to investigate risk of bias, study quality, and study limitations of meta-analyses of randomized controlled trials. *Two-thirds or more of studies judged to be low risk of bias = 0.5 points; over one-third of studies judged to be high risk of bias = 0 points; unclear risk of bias = 0.25 points; not assessed = 0 points. max., maximum.

Recently, the Grading of Recommendations Assessment, Development, and Evaluation (GRADE)[13] working group (4) developed a common and transparent approach for grading the quality of evidence and strength of recommendations based on systematic reviews, which was primarily developed for clinical guidelines (5, 6). The GRADE approach is now also being applied increasingly in nutrition research. In the field of nutrition research, several limitations arise when applying the GRADE criteria that should be considered, because systematic reviews of randomized controlled trials (RCTs) of nutrition interventions have inherent methodologic constraints. For example, RCTs of dietary interventions cannot be controlled with true placebos, but rather with certain constraints on nutrient compositions, food groups, or dietary patterns. Other limitations include lack of double blinding, poor compliance and adherence, crossover bias, and high dropout rates. Thus, in the field of nutritional epidemiology, in which RCTs are constrained, well-designed prospective cohort studies can provide important evidence (7).

The GRADE recommendation classifies systematic reviews of RCTs with an initial score of high and classifies systematic reviews of cohort studies with a score of low (8). To complement this methodologic gap, improved measures and tools that also take into account nutrition research–specific requirements (e.g., dietary assessment methods and their validation or funding bias) for assessing the meta-evidence (defined as the quality of evidence of meta-analyses: confidence in the estimate) need to be developed.

The aim of the current methodologic study was to design an improved grading approach (NutriGrade scoring system) tailored to nutrition research with the use of empirical data collected with previously developed tools and expert opinion.

NutriGrade aims to assess the meta-evidence of an association or effect between different nutrition factors and outcomes (e.g., hard clinical end points or surrogate markers). The newly developed scoring system was applied to a range of meta-analyses to test its performance and feasibility. In addition, we aimed to evaluate the performance of NutriGrade with the use of meta-analyses of diet–disease relations that already have applied the GRADE criteria.

Background of NutriGrade

The development of NutriGrade was based on a 3-stage approach: *1*) the planning phase, *2*) the design and development phase, and *3*) the validation phase.

As an initial basis, we categorized the single items of NutriGrade according to the GRADE approach: risk of bias, inconsistency, indirectness, publication bias, large effect, dose-response, and plausible confounding (4–6). Thereafter,

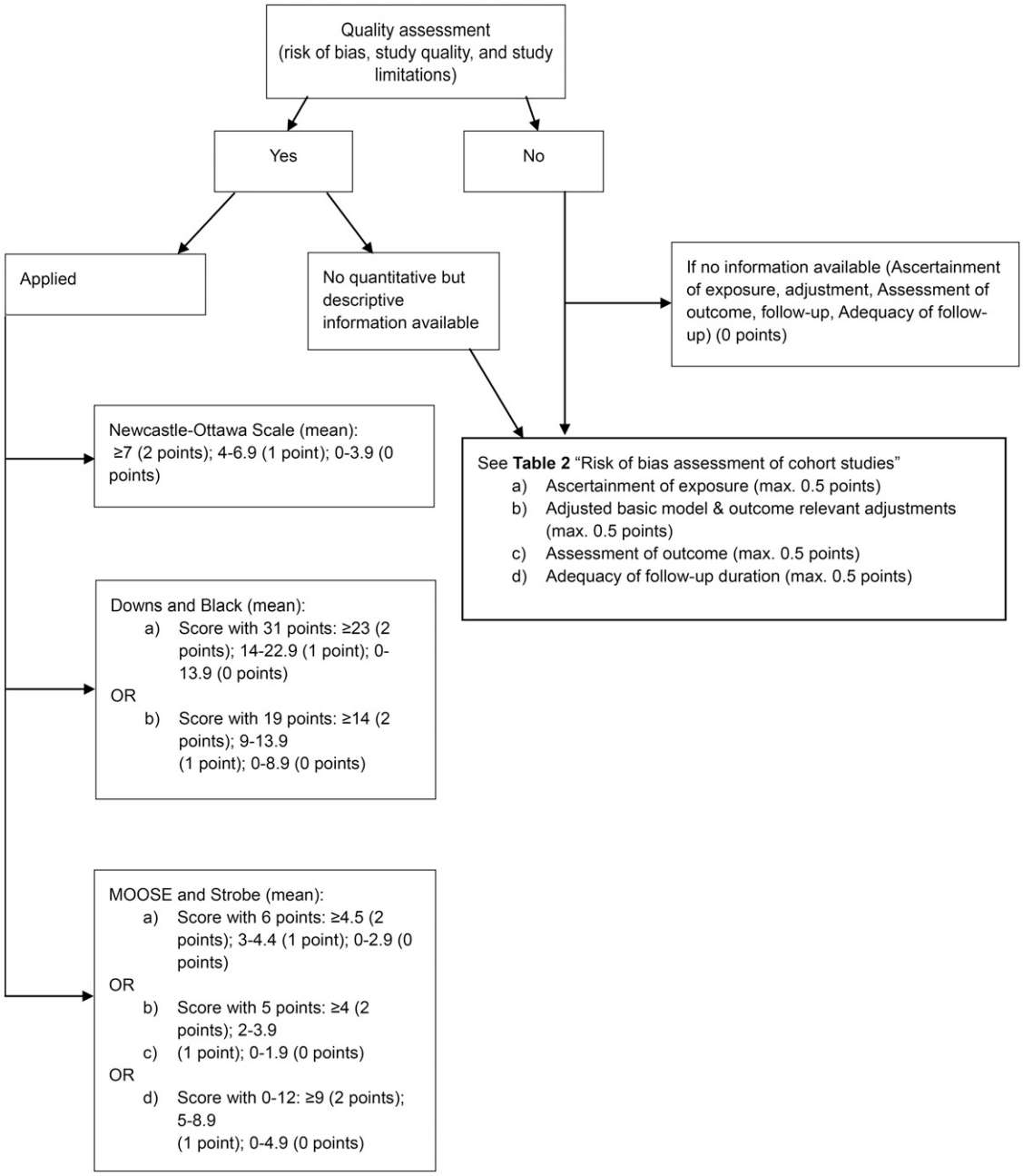

FIGURE 2 Procedure to investigate risk of bias, study quality, and study limitations of meta-analysis of cohort studies. max., maximum; MOOSE, Meta-analysis Of Observational Studies in Epidemiology. max., maximum.

TABLE 1 Risk of bias assessment of cohort studies (0–2 points)[1]

Subitems	Low risk of bias (two-thirds or more of included studies) = 0.5 points for each subitem	High risk of bias (over one-third of included studies) = 0 points for each subitem	Unclear risk of bias = 0.25 points for each subitem
Ascertainment of exposure	E.g., validated, calibrated FFQ or 24-h recall, diet history, or diet records (multiple days); Diet-associated biomarkers, e.g., 24-h urine	E.g., unvalidated FFQ, single 24-h recall, diet records, or diet history; Diet associated biomarkers: morning urine; Or not assessed	Assessed, but unclear[2]
Adjusted basic model and outcome-relevant adjustments	Basic model: ≥2 factors—sex, age, education, ethnicity; if only one sex included, then ≥1 factor; Outcome relevant adjustments: ≥3 factors—alcohol, energy intake, smoking, physical activity, BMI, CVD risk factors (blood pressure, dyslipidemia, family history of CVD)	Basic model: <2 factors—sex, age, education, ethnicity; if only one sex included, then ≤1 factor; Outcome-relevant adjustments: <3 factors—alcohol, energy intake, smoking, physical activity, BMI, CVD risk factors (blood pressure, dyslipidemia, family history of CVD); Or not assessed	Assessed, but unclear[2]
Assessment of outcome	E.g., record linkage (ICD codes), accepted clinical criteria, independent or blind assessment	E.g., self-report (not validated); Or not assessed	Assessed, but unclear[2]
Adequacy of follow-up duration	E.g., median ≥10 y for CVD, median ≥5 y for T2DM	E.g., median <10 y for CVD, median <5 y for T2DM; Or not assessed	Assessed, but unclear[2]

[1] CVD, cardiovascular disease; ICD, International Classification of Diseases; T2DM, type 2 diabetes mellitus.
[2] E.g., when too few details are available to make a judgment for low or high risk of bias.

we added the funding bias item and modified the items for nutrition research–specific requirements.

The rationale of the single grading system for each NutriGrade item and its detailed content was established after reviewing available guidelines on how to report systematic reviews: the Meta-analysis Of Observational Studies in Epidemiology checklist (9), the Cochrane Handbook (2), the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (10, 11), the Preferred Reporting Items for Systematic review and Meta-Analysis Protocols guidelines (12), the assessment tool for evaluating the methodological quality of systematic reviews (AMSTAR) (13), and the risk of bias in systematic reviews tool (14), systematic reviews on grading scores for observational studies, and RCTs (15, 16). After profound discussions between the authors, 9 items finally were selected for the NutriGrade scoring system. In each of the development stages, the scoring items were refined.

A detailed version with the rationale for the scoring of each item is available in the **Supplemental Appendix**.

NutriGrade Scoring System

The NutriGrade scoring system includes 7 items for meta-analyses of RCTs and 8 items for meta-analyses of cohort studies as follows: *1*) risk of bias, study quality, and study limitations, *2*) precision, *3*) heterogeneity, *4*) directness, *5*) publication bias, *6*) funding bias, *7*) study design (only for meta-analyses of RCTs), *8*) effect size, and *9*) dose-response (the latter 2 items for meta-analyses of cohort studies only; www.nutrigrade.net) (**Supplemental Tables 1 and 2**).

Risk of bias, study quality, and study limitations. Regarding RCTs, failure of allocation concealment, blinding, and follow-up losses are well-established limitations of RCTs (17, 18). Low-quality RCTs may lead to overestimation of intervention effect estimates and raise heterogeneity (19). Ascertainment of exposure, adjustment factors, assessment of outcome, and adequacy of follow-up are important limitations of cohort studies.

TABLE 2 Assessing precision in NutriGrade (0–1 points)

	0 points	1 point
Meta-analysis of RCTs[1]	*1*) <400 participants	*1*) 400–2000 participants, but 95% CI excludes the null value
	2) 400–2000 participants, but 95% CI includes the null value	*2*) >2000 participants
Meta-analysis of cohort studies	*1*) <500 events	*1*) ≥500 events and the 95% CI excludes the null value
	2) ≥500 events, but 95% CI includes the null value (i.e., CI includes RR of 1.0), and 95% CI fails to exclude important benefit (RR of <0.8) or harm (RR of >1.2)	*2*) ≥500 events, but 95% CI overlaps the null value (i.e., CI includes RR of 1.0), and 95% CI excludes important benefit (RR of <0.8) or harm (RR of <1.2)

[1] RCT, randomized controlled trial.

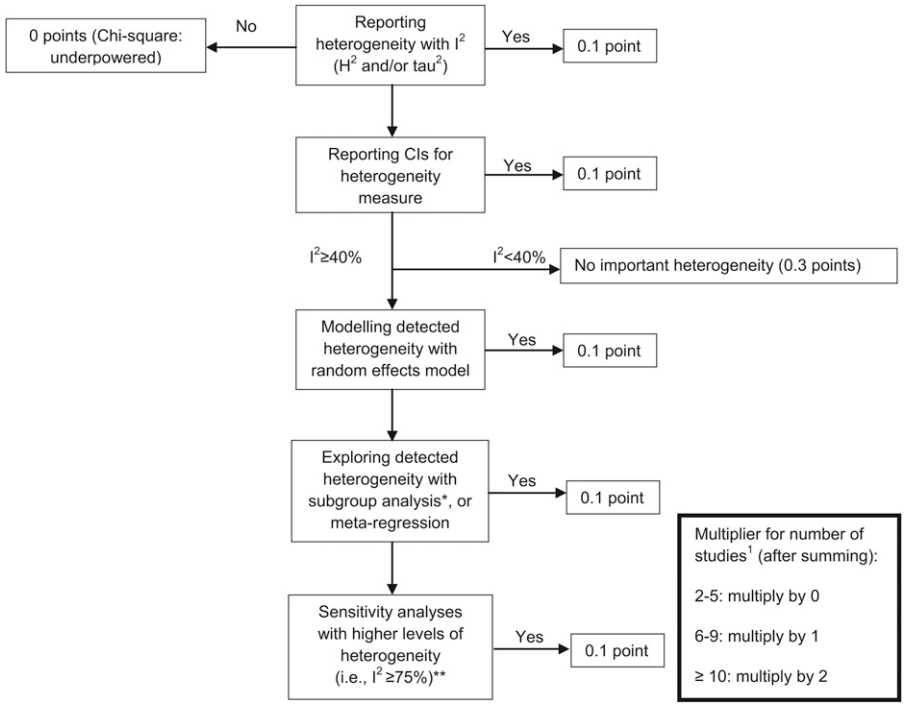

FIGURE 3 Assessing heterogeneity in meta-analyses. *Subgroup analyses: assessing whether effect is similar across specific groups of patients or is modified by study and patient characteristics (i.e., checking for consistency across subgroups, checking whether primary results are statistically significant). **Sensitivity analysis: repetition of the primary analysis (i.e., inclusion of some studies is unclear, because full details are not available; or exclusion of the largest study). For multiplier for number of studies, when authors treated men and women as separate studies, these should be treated as one study.

Risk of bias, study quality, and study limitations for meta-analyses of RCTs (maximum of 3 points). We included the risk of bias assessment tool by the Cochrane collaboration (20) for the NutriGrade scoring system. The judgments of the Cochrane risk of bias tool are expressed simply as low risk, high risk, or unclear risk of bias. We awarded each of the 6 risk-of-bias subitems (sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment personnel, incomplete outcome, and selective reporting) in a meta-analysis of intervention trials with 0.5 points (if two-thirds or more of the included studies were judged to be low risk of bias) or with 0 points (if over one-third of the included studies were at high risk of bias); otherwise (e.g., 80% of included studies were judged to be unclear risk), we awarded 0.25 points. Summing up the 6 subitems, each meta-analysis could be awarded a maximum of 3 points (**Figure 1**).

Risk of bias, study quality, and study limitations for meta-analyses of cohort studies (maximum of 2 points). We defined cutoffs for the identified quality scores (i.e., for the Newcastle–Ottawa Scale: ≥7 = 2 points; 4–6.9 = 1 point; 0–3.9 = 0 points) (**Figure 2**).

Because 40% of the identified meta-analyses of cohort studies applied no quality assessment, we developed a risk-of-bias checklist with 4 subitems (**Table 1**), awarding a maximum of 2 points (maximum of 0.5 points for each subitem). The risk of bias item should be applied if no quality assessment is reported.

Precision (maximum of 1 point). Statistical precision raises our confidence in the effect estimate. Precision previously was evaluated through the number of cases (events), sample size, and inspection of the 95% CIs. The number of points to be awarded for this item is given in **Table 2**.

Heterogeneity (maximum of 1 point). Checking consistency of the results is of major importance in meta-analyses. Statistical heterogeneity in studies is characterized by 95% CIs that show poor overlap. Methods to detect heterogeneity include the chi-square (Cochrane's Q) test and the I^2 statistic, among others (21, 22). If considerable heterogeneity is observed (I^2 ≥40%), it is recommended that possible reasons (e.g., differences between studies or data extraction errors)

TABLE 3 Assessment of directness (0–1 points)[1]

0 points (indirectness)	1 point
1) Differences in population: award 0 points only if there are important reasons to think that the physiology of the population of interest does not conform with that of the population tested, potentially leading to a considerably different effect measure	No important differences in the population or intervention; hard clinical outcome
2) Differences in intervention: applicable in RCTs; e.g., supervised vs. nonsupervised exercise; systematic difference in care between intervention and control group	
3) Surrogate markers (e.g., blood lipids, blood pressure)	
4) Network meta-analyses of RCTs; only a few available to date in nutrition research	

[1] RCT, randomized controlled trial.

TABLE 4 Assessment of publication bias (0–1 points)

0 points	0.5 points	1 point
1) <5 studies[1]	*1)* No evidence for publication bias with test or plot (5–9 studies)[1]	No evidence for publication bias with test or plot (≥10 studies)
2) Evidence for severe bias with test or plot	*2)* Evidence for moderate or small amount of publication bias with test or plot (≥10 studies)[1]	
3) Publication bias not assessed		

[1] When authors treated men and women as separate studies, here we count it as one study.

be explored (23, 24). The assessment procedure of heterogeneity in meta-analyses is shown in **Figure 3**.

Directness of evidence (maximum of 1 point). Direct evidence is characterized by similarities between the interventions or exposures of interest and populations of interest, and also includes outcomes important to the relevant patients and populations (e.g., stroke in a general population) (25). The scoring procedure for this item is shown in **Table 3**.

Publication bias (maximum of 1 point). Several investigations have shown that clinical trials with positive findings more often get published than those with nonsignificant findings (26, 27). The funnel plot (28), a graphic method, and a statistical test such as the Egger and Begg's are commonly used to detect publication bias in meta-analyses (29–31), although their interpretation is prone to errors (32, 33). Therefore, until now, there has been no gold standard to detect publication bias (2).

We propose the assessment for publication bias in the meta-analyses given in **Table 4**.

Funding bias (maximum of 1 point). Funding bias plays an important role in nutritional sciences (34, 35). Industry funding of nutrition-related scientific articles may bias conclusions in favor of sponsors' products, with potentially substantial implications for public health (36). The steps to assess funding bias are shown in **Table 5**.

Study design (for RCTs). The judgment of the quality of evidence recommended by the GRADE working group, in which observational studies (no differentiation between different types of observational studies) started with the low quality of a body of evidence, has several limitations in nutrition research (37–39) and should be modified or revised. According to the Oxford Centre for Evidence-Based Medicine, the level of evidence for RCTs is 1b, whereas cohort studies are graded as 2b (3). On the basis of these recommendations, we suggest awarding meta-analyses of RCTs with 2 points.

Effect size (only for cohort studies; maximum of 2 points). The definition of a meaningful effect, e.g., RR, HR, or OR, depends on the phenomena being studied. It is generally assumed that very large effects are less likely driven by confounding. The GRADE working group stated that a large effect can be assumed when having observed an RR of 2–5, or 0.5–0.2 (40). However, such large risk estimates often are not seen for nutrition and dietary exposures. The scoring procedure for effect size is given in **Table 6**.

Dose-response (for cohort studies; maximum of 1 point). Any type of dose-response gradient (linear and/or nonlinear) is an important factor for the presence of a causal relation (41). A dose-response association increases the confidence in the findings of cohort studies and thus enhances the assigned meta-evidence. The scoring for this item is explained in **Table 7**.

Agreement and Reliability

In a first step, 100 meta-analyses (**Supplemental References**) of prospective studies (observational and interventional) on foods and nutrients in relation to risk of chronic disease or cardiovascular disease risk factors were selected for the test. An independent researcher not involved in the project performed this selection with the use of electronic databases of meta-analyses (until 30 September 2015).

Five raters (BM, SD, CS, MS-M, and FE) with expertise in nutrition or epidemiology pretested the pilot version of NutriGrade. For this exercise, 6 meta-analyses were randomly selected from 100 meta-analyses. The raters were given guidance with regard to the interpretation of items included in the NutriGrade scoring system before reviewing the meta-analyses. GRADE recommends a maximum of 9 outcomes

TABLE 5 Assessing funding bias in meta-analyses (0–1 points)

0 points	0.5 points	1 point
Industry funding; OR Conflict of interest (1st, 2nd, or last author; e.g., member of advisory boards from the food industry, sale of books)	Private institutions, foundations, nongovernmental organizations (affiliation for each author should be checked)	Academic institutions, research institutions

TABLE 6 Scoring for effect size based on risk estimates (0–2 points)

0 points	1 point	2 points
No effect (RR or HR: 0.80–1.20) when comparing the highest vs. lowest category (e.g., in the mean, the comparison should be based on quartiles)	Moderate effect size (RR or HR: <0.80–0.50 and >1.20–2, and corresponding test is statistically significant) when comparing the highest vs. lowest category (e.g., in the mean, the comparison should be based on quartiles)	Large effect size (RR or HR: <0.50 and >2.00, and corresponding test statistically significant) when comparing the highest vs. lowest category (e.g., in the mean, the comparison should be based on quartiles)

to evaluate the quality of evidence, but we used only one outcome in this study to enable the raters to have a larger sample of different meta-analyses. Raters based their assessment on the information reported in each selected meta-analysis and did not seek information from the original reports of the included studies (RCTs and cohort studies). Each rater recorded the time spent conducting the assessment for each meta-analysis. From this data, we summarized the time spent to estimate the likely resource implications of using NutriGrade. Moreover, 2 authors (CS and FE) were randomly selected to repeat the exercise after 2 wk to analyze test–retest and intrarater agreement.

The reliability of ratings of NutriGrade items was calculated with the use of multirater κ values, and the reliability of the total (summed scores) was calculated with the use of intraclass correlation coefficients (ICCs). κ values were interpreted as suggested by Landis and Koch (42), and ICC values were interpreted with suggestions from Fleiss (43). Moreover, we calculated test–retest agreement by randomly selecting 2 independent raters (CS and FE) to repeat the assessment of 6 papers after a 2-wk interval. For the statistical analyses we used the statistical software package R with the R packages "irr" and "ICC" (44).

On the basis of the rating exercise described in the above pretest, agreement between raters was moderate to high. However, 3 of 9 items showed only fair agreement (κ = 0.21–0.40). To address this issue, we made some modifications and clarifications for the following items: risk of bias, study quality, and study limitations, heterogeneity, and funding bias.

Specifically, we included an additional figure (Figure 3) to facilitate heterogeneity assessment, we added more details (e.g., member of advisory boards or sale of books) on how to evaluate funding bias, and, for the risk of bias item, we added examples for low risk and high risk of bias studies.

After refining the NutriGrade scoring system, we used a random sample of 30 meta-analyses from the previously selected 100 meta-analyses while following a methodology similar to that reported by Shea et al. (13) for a pilot study. The included 30 meta-analyses covered a broad range of quality, albeit with some underrepresentation of high-quality meta-analyses. The same 5 raters (BM, SD, CS, MS-M, and FE) applied the new developed scoring system to all 30 meta-analyses.

TABLE 7 Scoring dose-response analysis (0–1 points)

0 points	1 point
No dose-response analysis or dose-response analysis with corresponding statistical test nonsignificant	Dose-response relation in prospective cohort studies: linear and/or nonlinear (corresponding statistical test significant)

The performance of NutriGrade improved considerably after revision of the pilot version. The meta-evidence score varied between 2.9 (very low meta-evidence) and 8.8 (high meta-evidence) for meta-analyses of intervention trials, and it ranged between 3.1 and 8.8 for meta-analyses of cohort studies. In eight meta-analyses, the maximum difference by the 5 raters differed by ~2 points.

The κ value of the ratings for each item ranged from 0.32 (95% CI: 0.22, 0.42) for risk of bias for cohort studies to 0.95 (95% CI: 0.91, 0.99) for study design, with a mean κ of 0.66 (95% CI: 0.53, 0.79). Moreover, item 5 (publication bias) scored "moderate" agreement, whereas all other items, with the exception of item 4 (directness) and item 7 (study design), scored "substantial" agreement (**Table 8**). The ICC of the total score for NutriGrade was 0.81 (95% CI: 0.69, 0.90), suggesting excellent reliability. A sensitivity analysis that excluded rater 3 showed a significant higher κ (0.45) for risk of bias of cohort studies.

Application of NutriGrade proved to be highly feasible for raters. The raters needed ~17 min to assess each paper (range: 7–42 min). Test–retest results performed by 2 raters showed high agreement (93% and 89%).

Judging the Meta-Evidence with NutriGrade

The present version of the NutriGrade scoring system is shown in **Table 9**. On the basis of this scoring system, we recommend 4 categories (45) to judge the meta-evidence: high, moderate, low, and very low, taking into account the following cutoffs: ≥8 points (high meta-evidence); 6–7.99 points (moderate meta-evidence); 4–5.99 points (low meta-evidence); and 0–3.99 points (very low meta-evidence) (**Table 10**).

Comparison with GRADE

Finally, we compared the GRADE quality of evidence judgment for 10 nutrition-related meta-analyses with the results of the NutriGrade application (which was performed independently by 2 raters) (5, 6, 46–53).

The application of the GRADE criteria to 5 meta-analyses of RCTs by choosing important or critical outcomes indicated a moderate quality of evidence for all meta-analyses tested. In contrast, with the application of NutriGrade, the meta-analyses scored low (3 meta-analyses) to moderate (2 meta-analyses) (**Table 11**). Applying the GRADE

TABLE 8 Assessment of multirater agreement for NutriGrade[1]

Item	(95% CI)
1) Risk of bias, study quality, and study limitations	RCTs Risk of bias: 0.70 (0.65, 0.75) Quality score: 0.67 (0.36, 0.98) Cohort studies Risk of bias: 0.32 (0.22, 0.42) Quality score: 0.67 (0.46, 0.88)
2) Precision	0.63 (0.52, 0.75)
3) Heterogeneity	0.64 (0.60, 0.68)
4) Directness	0.84 (0.72, 0.95)
5) Publication bias	0.50 (0.42, 0.59)
6) Funding bias	0.64 (0.54, 0.74)
7) Study design (only RCTs)	0.96 (0.91, 0.99)
8) Effect size (only cohort studies)	0.66 (0.50, 0.82)
9) Dose-response (only cohort studies)	0.76 (0.60, 0.92)

[1] RCT, randomized controlled trial.

criteria to meta-analyses of cohort studies resulted in very low- (2 meta-analyses) and low- (3 meta-analyses) quality evidence judgments. The NutriGrade application, on the other hand, indicated low (4 meta-analyses) and moderate (1 meta-analysis) meta-evidence for the same studies.

Discussion

A scoring system for meta-evidence in nutrition research was developed to address specific requirements and increasing needs to summarize the overwhelming amount of meta-analyses. This scoring system based on NutriGrade showed fair to substantial agreement for the different items and excellent agreement for the sum scores assigned to meta-analyses in nutrition research. In addition to having these appealing attributes, the implementation of NutriGrade was assessed to be feasible, with a mean of 17 min of review needed for each meta-analysis.

A special feature of NutriGrade is the inclusion of specific nutrition-relevant requirements, such as dietary assessment methods and their validation, calibration of FFQs, or the assessment of diet-associated biomarkers. Moreover, NutriGrade takes into account the limitation of blinding of participants and personnel in dietary intervention trials. Another important item that has been discussed as a limitation in previous assessment tools was not accounting for conflict of interest (54). As previous studies have shown, funding bias is of particular importance in nutrition research, with potential substantial public health consequences (36). In addition, the effect size item for evaluating meta-analyses of cohort studies has been adapted to more realistic cutoffs.

The most important benefit of NutriGrade, which, to our knowledge, is novel, is the modified classification for meta-analyses of RCTs and cohort studies compared with GRADE (which classifies systematic reviews of RCTs with an initial score of high, and systematic reviews of cohort studies with low). There has been a long debate regarding the best evidence in nutrition research, and whether it emerges from intervention trials in which the effects of a dietary change on disease, surrogate markers, or recognized risk markers are evaluated (55, 56). However, most dietary intervention trials are of short duration without targeting clinical outcomes such as morbidity or mortality. RCTs, if well-designed and -conducted, give robust answers to the research questions they address, but these do not include the investigations of long-term lifestyle behaviors on hard outcomes; conversely, observational data provide less-robust information regarding causality but usually address the question of directness of the study results. However, it is counterproductive to argue that, in general, one study design is superior to another (57). Based on the continuous scoring system of NutriGrade, we could show that meta-analyses previously evaluated with the GRADE tool reached a similar grading for meta-analyses of RCTs, but not for cohort studies. The assessment of publication bias and heterogeneity by taking into account the number of studies in NutriGrade might also help to adequately judge a meta-analysis based on cohort studies.

GRADE has been criticized recently for not providing sufficient guidance to make reliable and consistent judgments (58), e.g., low reliability for scoring quality of evidence domains (59). NutriGrade is based on clear guidance, with an assessment similar to a checklist. The main items, such as risk of bias, heterogeneity (inconsistency), precision, and directness, benefit by the stringent scoring system of NutriGrade, which showed better reliability than did GRADE in our study.

NutriGrade could be applied by authors of meta-analyses, who aim to summarize the meta-evidence for different outcomes. First, each chosen outcome would be assessed by the single items of NutriGrade. Second, the sum score of NutriGrade would be translated into meta-evidence

TABLE 9 Summary of Nutrigrade items

Item	RCTs[1] (maximum of 10 points)	Cohort studies (maximum of 10 points)
1) Risk of bias, study quality, and study limitations	0–3 points	0–2 points
2) Precision	0–1 point	0–1 point
3) Heterogeneity	0–1 point	0–1 point
4) Directness	0–1 point	0–1 point
5) Publication bias	0–1 point	0–1 point
6) Funding bias	0–1 point	0–1 point
7) Study design	+2 points	—
8) Effect size	—	0–2 points
9) Dose-response	—	0–1 point

[1] RCT, randomized controlled trial.

TABLE 10 Grading scoring system meta-evidence

NutriGrade score	Meta-evidence	Explanation
≥8 points	High	There is high confidence in the effect estimate, and further research probably will not change the confidence in the effect estimate.
6–7.99 points	Moderate	There is moderate confidence in the effect estimate; further research could add evidence on the confidence and may change the effect estimate.
4–5.99 points	Low	There is low confidence in the effect estimate; further research will provide important evidence on the confidence and likely change the effect estimate.
0–3.99 points	Very low	There is very low confidence in the effect estimate; meta-evidence is very limited and uncertain.

classes according to predefined cutoffs (very low, low, moderate, and high). A new research area in which NutriGrade could have a potential impact is the field of umbrella reviews (systematic reviews of systematic reviews and meta-analyses) and meta-epidemiologic studies. With NutriGrade, the overwhelming amount of meta-analyses for specific research questions could be summarized and allow for assessment of the overall meta-evidence, because meta-analyses on the same topic sometimes reveal different results (60, 61). A recent umbrella review summarizing the association between foods and nutrients and the risk of coronary heart disease, stroke, and type 2 diabetes included 93 meta-analyses (62). However, without applying the NutriGrade scoring system for each of these meta-analyses, the meta-evidence (the confidence in each estimate) for the associations between dietary exposure and outcome remains uncertain.

Although NutriGrade has been developed for nutrition research, it can be applied also to other lifestyle-related fields, such as physical activity, in which cohort studies add important evidence. In this case, authors should consider the validity of physical activity measurements.

The NutriGrade scoring system did not take into account the methodologic quality of systematic reviews. To assess the methodologic quality of systematic reviews and meta-analyses, a well-established tool (AMSTAR) already exists (13). Pollock et al. (58) recently developed an algorithm to assign GRADE levels of evidence that combines the assessment of methodologic quality of a systematic review with 3 items to

TABLE 11 Comparing GRADE quality of evidence with meta-evidence judgment based on the NutriGrade[1]

Comparison outcome reference	Study design	GRADE quality of evidence	NutriGrade	Meta-evidence judgment
Low saturated fat compared with usual diet Outcome: all-cause mortality (6)	MA of RCTs	Moderate	7.3	Moderate
Paleolithic diet Outcome: waist circumference (46)	MA of RCTs	Moderate	5.75	Low
High- vs. low-protein diet Outcome: weight loss (47)	MA of RCTs	Moderate	4.05	Low
High- vs. low-fat diet Outcome: TGs (48)	MA of RCTs	Moderate	6.8	Moderate
Low-carbohydrate vs. balanced diet Diastolic blood pressure (49)	MA of RCTs	Moderate	5.35	Low
High- vs. low-saturated fat Outcome: all-cause mortality (5)	MA of cohort studies	Very low	4.5	Low
High vs. low coffee intake Outcome: endometrial cancer (50)	MA of cohort studies	Low	7.3	Moderate
High vs. low sodium intake Outcome: stroke (51)	MA of cohort studies	Very low	5.55	Low
High vs. low intake of sugar-sweetened beverages in children Outcome: adiposity (52)	MA of cohort studies	Low	4.5	Low
High vs. low potassium intake Outcome: stroke (53)	MA of cohort studies	Low	5.9	Low

[1] GRADE, Grading of Recommendations Assessment, Development and Evaluation; MA, meta-analysis; RCT, randomized controlled trial.

grade the quality of evidence (risk of bias of trials, heterogeneity, and number of participants). However, this approach has been criticized by members of the GRADE working group as problematic and misleading (and should not be part of the GRADE scoring system as an additional factor that influences the quality of the body of evidence). These authors suggested that systematic reviews of low quality should be excluded from umbrella reviews (63). The Dietary Guidelines for Americans 2015 included only systematic reviews and meta-analyses of a methodologic quality score of ≥8 (out of 11, according to AMSTAR) in its recent report (64). However, it could be shown that the methodologic quality of systematic reviews investigating the effect of a Mediterranean diet on cardiovascular disease did not fully meet the standards (65). Therefore, when applying the NutriGrade scoring system, we recommend assessing the methodologic quality of meta-analyses by also applying the AMSTAR tool.

Studies testing the reliability of the risk of bias assessment tool by the Cochrane collaboration and Newcastle–Ottawa Scale showed that more detailed guidance would be useful, because the agreement between rates was low (66–68). Moreover, Stang (69) commented that the Newcastle–Ottawa Scale includes quality items that are not valid (e.g., the "representativeness of the exposed cohort" item), and concluded that this score appeared to be unacceptable for the quality ranking of case–control and cohort studies in meta-analyses. Therefore, we recommend that our risk of bias assessment tool be applied to ongoing meta-analyses of cohort studies.

A meta-analysis provides important insights for the judgment of the strength of association and the dose-response for the relation of a nutrition factor and an outcome. The identified strength of association commonly termed as effect size and the dose-response alone are not sufficient for causal claims (70). However, a meta-analysis provides a basis to figure out different sources of confounding and bias from inclusion of different studies with different study characteristics, thereby providing useful information for further causal research.

Conclusion

So far, NutriGrade has shown good agreement between raters and convincing reliability. The reported findings regarding the performance of the scoring system need further confirmation by a broader range of applications with more reviews evaluated with NutriGrade.

Acknowledgments

We thank Martin Wiseman for his many helpful comments on an earlier version of this manuscript. All authors read and approved the final manuscript.

References

1. Mann JI. Evidence-based nutrition: does it differ from evidence-based medicine? Ann Med 2010;42:475–86.
2. Higgins JPT, Green S (editors) [Internet]. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011 [cited 2015 Dec 8]. Available from: www.cochrane-handbook.org.
3. Oxford Centre for Evidence-Based Medicine [Internet]. Levels of evidence [cited 2015 Nov 5]. Available from: http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/.
4. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.
5. de Souza RJ, Mente A, Maroleanu A, Cozma AI, Ha V, Kishibe T, Uleryk E, Budylowski P, Schunemann H, Beyene J, et al. Intake of saturated and trans unsaturated fatty acids and risk of all-cause mortality, cardiovascular disease, and type 2 diabetes: systematic review and meta-analysis of observational studies. BMJ 2015;351:h3978.
6. Hooper L, Martin N, Abdelhamid A, Davey Smith G. Reduction in saturated fat intake for cardiovascular disease. Cochrane Database Syst Rev 2015;6:CD011737.
7. Satija A, Yu E, Willett WC, Hu FB. Understanding nutritional epidemiology and its role in policy. Adv Nutr 2015;6:5–18.
8. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401–6.
9. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283:2008–12.
10. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151:W65–94.
11. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151(4):W65–94.
12. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1.
13. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.
14. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225–34.
15. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666–76.
16. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62–73.
17. Schwingshackl L, Missbach B, Dias S, Konig J, Hoffmann G. Impact of different training modalities on glycaemic control and blood lipids in patients with type 2 diabetes: a systematic review and network meta-analysis. Diabetologia 2014;57:1789–97.
18. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. Cochrane Database Syst Rev 2012;3:CD007176.
19. Savović J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, Als-Nielsen B, Balk EM, Gluud C, Gluud LL, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012;157:429–38.
20. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
21. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
22. Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. Stat Med 2006;25:4321–33.
23. Thompson SG, Turner RM, Warn DE. Multilevel models for meta-analysis, and their application to absolute risk differences. Stat Methods Med Res 2001;10:375–92.

24. Gøtzsche PC, Hrobjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007;298:430–7.
25. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, et al. GRADE guidelines: 8. Rating the quality of evidence–indirectness. J Clin Epidemiol 2011;64:1303–10.
26. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Assess 2000;4:1–115.
27. Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev 2007; (2):MR000005.
28. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001;54:1046–55.
29. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med 2006;25:3443–57.
30. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA 2006;295:676–80.
31. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34.
32. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006;333:597–600.
33. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol 2005;58:894–901.
34. Nestle M. Food company sponsorship of nutrition research and professional activities: a conflict of interest? Public Health Nutr 2001;4:1015–22.
35. Rowe S, Alexander N, Clydesdale FM, Applebaum RS, Atkinson S, Black RM, Dwyer JT, Hentges E, Higley NA, Lefevre M, et al. Funding food science and nutrition research: financial conflicts and scientific integrity. Am J Clin Nutr 2009;89:1285–91.
36. Lesser LI, Ebbeling CB, Goozner M, Wypij D, Ludwig DS. Relationship between funding source and conclusion among nutrition-related scientific articles. PLoS Med 2007;4:e5.
37. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71–2.
38. Hernán MA, Hernandez-Diaz S, Robins JM. Randomized trials analyzed as observational studies. Ann Intern Med 2013;159:560–2.
39. Kroke A, Boeing H, Rossnagel K, Willich SN. History of the concept of 'levels of evidence' and their current status in relation to primary prevention through lifestyle interventions. Public Health Nutr 2004;7:279–84.
40. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311–6.
41. Hill AB. The environment and disease: association or causation? 1965. J R Soc Med 2015;108:32–7.
42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
43. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons Inc.; 1986.
44. R-package [Internet]. [cited 2016 Feb 1]. Available from: https://cran.r-project.org/web/packages/ICC/index.html; https://cran.r-project.org/web/packages/irr/irr.pdf.
45. Mente A, de Koning L, Shannon HS, Anand SS. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. Arch Intern Med 2009;169:659–69.
46. Manheimer EW, van Zuuren EJ, Fedorowicz Z, Pijl H. Paleolithic nutrition for metabolic syndrome: systematic review and meta-analysis. Am J Clin Nutr 2015;102:922–32.
47. Santesso N, Akl EA, Bianchi M, Mente A, Mustafa R, Heels-Ansdell D, Schunemann HJ. Effects of higher- versus lower-protein diets on health outcomes: a systematic review and meta-analysis. Eur J Clin Nutr 2012;66:780–8.
48. Schwingshackl L, Hoffmann G. Comparison of the long-term effects of high-fat v. low-fat diet consumption on cardiometabolic risk factors in subjects with abnormal glucose metabolism: a systematic review and meta-analysis. Br J Nutr 2014;111:2047–58.
49. Naude CE, Schoonees A, Senekal M, Young T, Garner P, Volmink J. Low carbohydrate versus isoenergetic balanced diets for reducing weight and cardiovascular risk: a systematic review and meta-analysis. PLoS One 2014;9:e100652.
50. Zhou Q, Luo ML, Li H, Li M, Zhou JG. Coffee consumption and risk of endometrial cancer: a dose-response meta-analysis of prospective cohort studies. Sci Rep 2015;5:13410.
51. Aburto NJ, Ziolkovska A, Hooper L, Elliott P, Cappuccio FP, Meerpohl JJ. Effect of lower sodium intake on health: systematic review and meta-analyses. BMJ 2013;346:f1326.
52. Te Morenga L, Mallard S, Mann J. Dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. BMJ 2013;346:e7492.
53. Aburto NJ, Hanson S, Gutierrez H, Hooper L, Elliott P, Cappuccio FP. Effect of increased potassium intake on cardiovascular risk factors and disease: systematic review and meta-analyses. BMJ 2013;346:f1378.
54. Lucas M. Conflicts of interest in nutritional sciences: The forgotten bias in meta-analysis. World J Methodol 2015;5:175–8.
55. Maki KC, Slavin JL, Rains TM, Kris-Etherton PM. Limitations of observational evidence: implications for evidence-based dietary recommendations. Adv Nutr 2014;5:7–15.
56. Ankarfeldt MZ. Comment on "Limitations of observational evidence: implications for evidence-based dietary recommendations". Adv Nutr 2014;5:293.
57. Dreyer NA, Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why observational studies should be among the tools used in comparative effectiveness research. Health Aff (Millwood) 2010;29:1818–25.
58. Pollock A, Farmer SE, Brady MC, Langhorne P, Mead GE, Mehrholz J, van Wijck F, Wiffen PJ. An algorithm was developed to assign GRADE levels of evidence to comparisons within systematic reviews. J Clin Epidemiol 2016;70:106–10.
59. Berkman ND, Lohr KN, Morgan LC, Kuo TM, Morton SC. Interrater reliability of grading strength of evidence varies with the complexity of the evidence in systematic reviews. J Clin Epidemiol 2013;66(10):1105–1117e1.
60. Schwingshackl L, Missbach B, Hoffmann G. An umbrella review of garlic intake and risk of cardiovascular disease. Phytomedicine 2015 Nov 14 (Epub ahead of print; DOI: 10.1016/j.phymed.2015.10.015).
61. Schwingshackl L, Hoffmann G. Monounsaturated fatty acids and risk of cardiovascular disease: synopsis of the evidence available from systematic reviews and meta-analyses. Nutrients 2012;4:1989–2007.
62. Mozaffarian D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity: A Comprehensive Review. Circulation 2016;133:187–225.
63. Murad MH, Mustafa R, Morgan R, Sultan S, Falck-Ytter Y, Dahm P. Rating the quality of evidence is by necessity a matter of judgment. J Clin Epidemiol 2016 Nov 14 (Epub ahead of print; DOI: 10.1016/j.phymed.2015.10.015).
64. The 2015 Dietary Guidelines Advisory Committee [Internet]. In Scientific report of the 2015 Dietary Guidelines Advisory Committee [cited 2016 Mar 2]. Available from: https://health.gov/dietaryguidelines/2015-scientific-report/PDFs/Scientific-Report-of-the-2015-Dietary-Guidelines-Advisory-Committee.pdf.
65. Huedo-Medina TB, Garcia M, Bihuniak JD, Kenny A, Kerstetter J. Methodologic quality of meta-analyses and systematic reviews on the Mediterranean diet and cardiovascular disease outcomes: a review. Am J Clin Nutr 2016;103:841–50.
66. Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, Dryden DM. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 2013;66:982–93.
67. Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, Tsertsvadze A, Hempel S, Shekelle P, Dryden DM. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;66:973–81.
68. Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, Fuentes J, Ha C, Cummings GG. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. PLoS One 2014;9:e96920.
69. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25:603–5.
70. Weed DL. Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int J Epidemiol 2000;29:387–90.